

How the online social networks are used: Dialogs-based structure of MySpace

Milovan Šuvakov^{1,2}, Marija Mitrović¹, Vladimir Gligorijević¹ and Bosiljka Tadić¹

¹Department of theoretical physics; Jožef Stefan Institute; Box 3000 SI-1001 Ljubljana Slovenia; ²Institute of Physics, Belgrade University, Belgrade, Serbia

Abstract: Quantitative study of collective dynamics in online social networks is a new challenge based on the abundance of empirical data. Conclusions, however, may depend on factors as user's psychology profiles and their reasons to use the online contacts. In this paper we have compiled and analysed two datasets from MySpace. The data contain networked dialogs occurring within a specified time depth, high temporal resolution, and texts of messages, in which the emotion valence is assessed by using SentiStrength classifier. Performing a comprehensive analysis we obtain three groups of results: Dynamic topology of the dialogs-based networks have characteristic structure with Zipf's distribution of communities, low link reciprocity, and disassortative correlations. Overlaps supporting "weak-ties" hypothesis are found to follow the laws recently conjectured for online games; Long-range temporal correlations and persistent fluctuations occur in the time series of messages carrying positive (negative) emotion; Patterns of user communications have dominant positive emotion (attractiveness) and strong impact of circadian cycles and interactivity times longer than one day. Taken together, these results give a new insight into functioning of the online social networks and unveil importance of the amount of information and emotion that is communicated along the social links. **All data used in this study are fully anonymized.**

Keywords: Online social networks; User communities; Patterns of emotion flow; Self-organized dynamics;

1 Introduction

Web as a new social space provides "unbearable easiness of communication" that may lead to new social phenomena in the online world and affect behavior of the users [1, 2, 3]. The online social networks nowadays represent some of the largest social structures in the world [4]. Apart from the structure, huge amount of users leave digital data about their activity, which are systematically stored at Web portals, thus providing "social experiments" of an unprecedented scale. Recently analysis of such data from various Web portals has been performed across different science disciplines [5, 6, 7, 8, 9, 10, 11, 12]. It has been recognized that the emotions, known to drive social behavior in face-to-face contacts, also play an important role in various types of the online interactions [13, 12, 14, 15, 16, 17, 18, 19]. However, full understanding of the mechanisms that drive online social behaviors still remains elusive. In particular, how individual (emotional) actions of users in the network may lead to dynamically robust collective behaviors and how to quantify the emergent phenomena, are among the questions of primary importance. What are characteristic features of the dynamics in the online social networks in comparison with Blogs, Forums, Games and other forms of online communications? Here we address these questions by compiling and analysing the empirical data of *user dialogs* from the social network MySpace.

In recent psychology research quantitative study of emotion [20, 21] is conducted based on Russell's multidimensional model of affect [22]. Specifically, each emotion known in common life can be represented by a set of numerical values in the corresponding multidimensional space [22]. Three important dimensions of emotion—valence, arousal and dominance, can be estimated from psycho-physiological and neuronal activity [23]. The *valence* measures degree of attractiveness (positive valence) or aversiveness (negative valence) to a stimuli. Similarly, the arousal and the dominance can have a range of values corresponding to different degree of reactivity to a stimuli, and power of a reaction, respectively. Moreover, normative emotional rating has been developed for a large number of words [24]. Based on the lexicon of emotional words and using machine-learning approaches recently methods have been developed [23, 16, 25] for the effective inference of the emotion content from different types of text messages appearing in the online communications.

How the online social networks are used and who uses them? The social psychology research begin to recognize the relationship between the personal profile, social and emotional loneliness and attitudes of the Web users in general, and users of the social networks, in particular [26, 27, 9, 28]. The "friendship" association as the framework for communications in the online social networks, is partially transferred from offline social contacts. Recently the topology of *friendship network* in currently largest online social site Facebook has been analysed

[4, 29]. Conclusion is that “mostly, but not entirely, agreement on common structural network characteristics” was found [4]. Somewhat different structure was reported for the *friendship network* of MySpace [30]. However, it is not mere structure of the network of declared “friendship” links, but rather the *dynamics of message exchange* along these links that contain relevant information for study the emergent social phenomena, in particular in the situations when temporal bursts of (emotional) messages occur involving many users. To our knowledge, such dynamical structures, reflecting the way how the links in the online social networks are used, have not been researched so far. Therefore, developing a methodology for systematic collection and analysis of the data which contain complete information about the dynamics of dialogs is of key importance for diagnostics of bursting events, detecting and characterizing collective behaviors, and identifying involved users.

In this work we study the dialogs-based social networks which represent dynamical structures situated on the underlying friendship network in MySpace. They involve only certain number (or type?) of the active users and their structure can vary in time, depending on the events and time window of interest. We consider two sets of data with the dialogs among users in MySpace social network and analyse them as complex dynamical systems in order to determine quantitative measures of users collective behaviors. For this purpose we first develop a procedure to compile the data that have required structure and high temporal resolution for this type of analysis. The datasets are then studied by methods of graph theory and statistical physics to determine topology of the dialogs-based networks and to define and compute several other quantitative measures of the collective behaviors, in particular, the temporal correlations and the patterns of user’s activity, that can be inferred from such data. Furthermore, using the emotion classifier SentiStrength [15, 16], which is developed for graded estimate of the emotional content in short informal texts, we assess the emotion valence in the text of each message in our datasets. This enables us to further analyse how the flow of emotion adheres with the dialogs-based network topology and with the observed collective behaviors of users.

2 Empirical data from MySpace social network

For the goals that we pursue in this paper high resolution data are necessary, in particular, (a) a *connected network* of users identified by their IDs as nodes, and the exchanged messages as links, and (b) *each message identified* by its source and target nodes (as user IDs), time when the message is generated, and its textual content. Arguably, an ideal online social network for this analysis is MySpace, having the data of such structure systematically recorded until 2010.

2.1 Data crawling & structure

Communications between users in the online social networks as MySpace occur along friendship links: writing messages on own *wall*, where they can be seen by the linked friends, or writing to the linked friend’s wall directly. Privacy policy vary from user to user, hence some users do not allow access to their data. However, the messages that they sent to the linked users who allow the access, can be identified.

To obtain the networked dialogs in MySpace, we developed a parametrized Web crawler who visits mutually linked users and collects the dialogs (messages) which were posted by them *within a specified time window*. To start we first specify the time window that we are interested in, and find an appropriate node, representing a *user* who was active within that time window. The crawler then proceeds to search in the neighborhood of that node in a *breadth-first manner*, as schematically shown in Fig. 1, and collecting all messages posted within that time window on the current user’s wall and identifying their source and posting time. The links in Fig. 1 indicate that at least one message were exchanged between the two users within the specified time window. Starting from the initial node marked as “1” the list of first layer of the connected nodes is explored, then the search is continued from each node on that list, thus making the second layer nodes. Then the lists are swapped, and so on. The crawler is instructed to avoid the nodes whose data are marked as “private” and also the nodes which contain too many connections (probably representing a non-human user). The crawler can identify the messages posted from these types of nodes, that can be seen at the walls of their neighbors. Such nodes and their messages are not considered in the analysis. Full information about the exchanged messages along each discovered link is stored in a database, in particular, the identity of the source and the target node, the creation time and full text of each

message. Given the parametrized search, the crawler can stop either when no new nodes are found which satisfy the parameter criteria (time depth) or when it reaches a given diameter (number of layers), or accidentally for some unexpected reason.

Our data [31] are compiled in 2009 and contain messages for two time windows—two months and three months depth, but starting from the same initial node. One dataset for two months period, January and February 2008, consists of $N_M = 80754$ messages exchanged between $N_U = 36462$ users. The larger dataset corresponds to time depth of three months, from 1st of June till 1st of September 2008, and contains $N_U = 64739$ users and $N_M = 172127$ of their messages.

2.2 Inference of the emotional valence from text of messages

For further analysis in this work we performed automatic screening of the text of dialogs in our dataset to extract the emotional valence of each message. We apply the Emotion classifier which is developed by Thelwall and coworkers [15, 16, 25] for short text messages, which occur in MySpace dialogs. The classifier considers each message as a single document and can extract graded emotion valence as two numbers (e_- , e_+), representing the intensity of the negative and the positive emotion content of the same message.

According to Refs. [16, 15, 25], the classifier algorithm is based on two emotional dictionaries *the General Inquirer* (GI) and *Linguistic Inquiry and Word Count* (LIWC). In each message the algorithm detects all words that belong to the emotional dictionary and extracts their polarity and intensity. The obtained scores are then modified with additional linguistic-based rules if special terms, such as negators (*good* vs. *bad*), intensifiers (*liked* vs. *liked very much*), diminishers (*excellent* vs. *rather excellent*) are found in the neighborhood of that word in the area of 5 words before and after the emotional term or the beginning or end of the sentence. Other special terms as capitalization (*bad* vs. *BAD*), exclamation and emotion detection (*happy!* or *: -)* are searched and treated similarly as intensifiers. If an intensifier (or diminisher) word is found then the absolute original emotional value of the term is increased (or decreased) by one. For example, *bad* is given (-3) then *very bad* is (-4) . Similarly, if *good* is judged as $(+3)$ then *somewhat good* is $(+2)$, while *very good* is $(+4)$. The scores e_+ and e_- that the classifier returns represent the maximum positive and the maximum negative number for a given message. Accuracy and other relevant details can be found in the original Refs. [16, 15, 25].

The classifier returns two independent ratings for every message: one for the positive $e_+ \in \{+1, +2, +3, +4, +5\}$ and one for the negative $e_- \in \{-1, -2, -3, -4, -5\}$ dimension. On this scale $e_- = -5$ correspond to very negative and $e_+ = +5$ to very positive emotion valence. While $e_- = -1$ and $e_+ = +1$ indicate the absence of negative and positive emotion, respectively. Based on these automated ratings we can construct the overall valence polarity of a particular message. Specifically, all messages for which the scores are in the range ($e_- \leq -3$ and $e_+ = +1$ or $e_+ = +2$) are considered as carrying *negative emotion valence*, and symmetrically the messages with ($e_+ \geq +3$ and $e_- = -1$ or $e_- = -2$) are classified as carrying *positive emotion valence*. Notice that this excludes certain number of messages for which the two scores are exactly balanced, ($e_- = e_+$), although they may contain “emotional” words. Also the messages for which the negative and positive scores are simultaneously very high, ($e_+ \geq 3$ and $e_- \leq -3$), are disregarded as possible artifact of the graded-strength classifier.

2.3 Methodology for data analysis

The data set for a given time window is mapped onto dialogs-based network in the following way: Each user is represented by a node and a link is inserted between the pair of nodes (i, j) if a dialog (at least one message) was detected between this pair of users within that time window. The direction of the link indicates the message from source-to-target and the link multiplicity (weight) represents the number of messages. In some figures colors of the links are used to suggest net emotion balance along the link within the specified time window, which is derived from the original scores of the messages along that link.

Our high-resolution datasets contain valuable information which enables us to study other features, in particular the temporal sequences of user activities and the patterns of emotion flow via message exchange. To study such diverse aspects of the online social networks, appropriate methodologies and mathematical approaches are

applied. In particular:

- *The network topology* is analysed using the graph theory methods [32]. Specifically, we determine several topology measures at global network level (distributions of degree, strength, weight, betweenness) and at local node-neighborhood level (assortativity measures and tests of social weak-tie hypothesis), as well as the mesoscopic level (community structure). For the community structure analysis in compact and relatively small networks we apply the eigenvalue spectral analysis of the Laplacian operator related with the *weighted symmetrical* adjacency matrix [33]. In the case of large graphs the communities are detected by Gephi software, which utilizes maximum modularity approach [34]. The maximum-flow spanning trees of our networks are determined using greedy algorithm.
- *Temporal correlations* are studied by time series analysis. Three types of time series are extracted from the data, i.e., the time series of the number of messages, and the number of messages carrying positive or negative emotion, per small time bin $t_{bin} = 5$ minutes.
- *Patterns of user activity* are mapped directly from the dataset in the original temporal resolution. Each action (message) of a given user at a given time is represented by a point on the user temporal pattern. The interactivity time is identified as the distance between two consecutive points along the time axis from the activity pattern of each user.

Various histograms obtained in this analysis are fitted to the corresponding theoretical expressions using either χ^2 or Maximum Likelihood Estimator (MLE) method. All fits passed the χ^2 -test of goodness. Logarithmic binning with the base $b = 1.1$ is often applied to obtain smooth curves.

3 Topology of the Dialogs-Based Networks in MySpace

Using the above described procedure of data mapping, we obtain networks of MySpace users, as nodes, connected by directed weighted links, representing the messages sent from one user to another. The weight of a link indicates the number of messages along that link within the considered time window. In Figure 2 two examples of such dialog-based networks are shown. In the top panel a part of the network of dialogs observed within the time window $T_{WIN} = 2$ months is shown. While the lower panel shows the corresponding network for the situation when the searched time depth is $T_{WIN} = 3$ months starting from the same initial node, as explained in section 2.1. Shown are small initial part of the corresponding dataset. The increased time depth manifests in that larger number of nodes are connected to the network, the links density is increased as well as the widths of some already existing links. Moreover, the community structure—visually identified as groups of nodes, is evolving.

Density and Reciprocity of links. Here we consider topology of two networks representing all *emotion classified dialogs* in 2 month time window (Net2M), and in 3 months time window (Net3M). They contain $N = 33649$ and 58957 users, respectively. These networks appear to be very sparse, cf. Fig. 3. The average link density, defined as $\rho \equiv L/N(N-1)$, where L is the number of all directed links, is found as $\rho = 3.345 \times 10^{-5}$ for Net2M, and $\rho = 2.08 \times 10^{-5}$ for Net3M network. Furthermore, we compute the *link reciprocity*, which is defined [35] as $r \equiv (L^{\leftrightarrow}/L - \rho)/(1 - \rho)$, where L^{\leftrightarrow} is the number of links occurring in both directions $i \rightarrow j$ and $j \rightarrow i$ disregarding the weight. We find $r = 0.0227$ for Net2M, and $r = 0.0214$ for Net3M network, i.e., the reciprocity is barely positive. According to [35], in social networks larger positive reciprocity is expected. The clustering coefficient $C_c = 0.013$ and 0.014 are found for these two networks when the directedness of the links is ignored. We also consider a reduced network, termed Net3321, which is extracted from two months dataset. In this network the users who sent and received less than four messages within two months were excluded. Thus reduced network contains $N_U = 3321$ nodes and is more compact, i.e., $\rho = 7.19 \times 10^{-4}$ and has larger link reciprocity $r = 0.118$ and the clustering coefficient $C_c = 0.084$, for directed, and 0.115 for undirected links.

Community structure. In Fig. 3 the entire network of Net2M is shown. The network is organized in large number of small *communities*, specifically 87 communities shown in Fig. 3 are obtained by weighted maximum modularity algorithm [34] with Gephi software. Largest community contains 2543 users. It is interesting to note that the size distribution of these communities obeys Zipf's law. In Fig. 4a two curves in the inset are the ranking distributions

of the community sizes which are detected in the networks of two months dialogs and three months dialogs, respectively. It should be stressed that the number of communities that one observes depends on the resolution in the algorithm. By increasing the resolution, some communities can further split into two or more groups, and oppositely, join together to make a larger community when the resolution is decreased. However, the scale-free organization of communities up to certain size persists (with a changed slope), as shown in the main Fig. 4a.

In the Net3321, the nodes with small strength $\ell_{in} + \ell_{out} \leq 4$ are excluded, as mentioned above. (The directed network is further analysed in section 4.) Here in Fig. 4b we confirm that this more compact network also exhibits a community structure. We use the eigenvalues spectral analysis [33] of the normalized Laplacian $\mathcal{L} = \delta_{ij} - \frac{w_{ij}^S}{\sqrt{\ell_i \ell_j}}$, where ℓ_i and ℓ_j are total strengths and $W_{ij}^S \equiv W_{ij} + W_{ji}$ symmetrical weighted adjacency matrix of the network. Property that the eigenvectors localize on subgraphs [33] is used to identify communities. Note that the communities detected in the dialogs-based network are dynamical, i.e., related with a particular part of the network structure that has been actually used within the considered time window.

Nodes inhomogeneity & mixing patterns. The node's degree and strength distributions and mixing patterns for the networks Net2M and Net3M are computed and the results are shown in Figs. 5 and 6. Both degree and strengths distributions can be fitted by the mathematical expression

$$P_{\kappa}(X) = B_{\kappa} X^{-\tau_{\kappa}} e^{-X/X_{0\kappa}} \quad (1)$$

where the exponent τ_{κ} and the characteristic cut-off length $X_{0\kappa}$ may vary, depending on the type of the links ($\kappa = \text{in, out}$) and the time-window where the dialogs are observed. Specifically, the distributions of out-link degree and out-link strengths follow a power-law decay. The exponent $\tau_{out} = 3.01 \pm 0.07$ before the cut-off $X_0 = 18$ for the degree, and $\tau_{out} = 1.62 \pm 0.02$ and $X_0 = 98$ for the strength distributions are found. In Fig. 5a,b fitted are only parts with the power-law dependence. These distributions appear to be quite stable with respect to time-depth, i.e., similar slopes are found for the distributions from 2-months and 3-months time window data. On the other hand, the in-links degree and in-links strengths receive the form (1) only when the time-depth is large enough. They are characterized by much smaller exponent $\tau_{in} < 1$ and large cut-off: $\tau_{in} = 0.53 \pm 0.07$, $X_0 = 226 \pm 32$ for the in-degree, and $\tau_{in} = 0.88 \pm 0.04$, $X_0 = 248 \pm 17$ for the in-strength. The best fit, as shown in Fig. 5, is received when a small stretching exponent is added, 1.25 for the in-degree and 1.06 for the in-strength distributions. While the out-degree and out-strength are controlled by the node itself—representing the user's actions directed to its neighbors, the in-degree contains the cumulative action of all neighbors directed to that node's wall. Hence, the lower exponent is expected as well as the large cut-off, which reflects the diversity of first-neighborhoods of nodes on the network.

The assortativity measures are another characteristics of the network's topology at the level of node's neighborhood. Specifically, the situations when the nodes with large degree are linked to each other (assortativity) or oppositely, the nodes with large degree have a large number of nodes with small degree (dis-assortativity), will be expressed by increasing (decreasing) slopes in the degree-correlations plots, like the ones in Fig. 6a. In view of the *weighted and directed* nature of the dialogs network, we can determine several such measures. The results are shown in Fig. 6a,b for degree and for strength mixing, respectively. Specifically, considering a node with a given in-degree, plotted along the x-axis $q_{\kappa=\text{in}}$, and computing the average out-degree of the nodes linked to it, $\langle q_{\mu=\text{out}} \rangle_{nn}$, referred as *in-out* in the Legend, we find a dis-assortative pattern, i.e., $\langle q_{\mu=\text{out}} \rangle_{nn} \sim q_{in}^{-\mu}$ with the decreasing slope $\mu \sim 1$. The tendency to flattening in the neighborhood of large-degree nodes suggests the occurrence of communities. Similar finding applies for the *out-in* mixing. On the other hand, no assortativity measures can be observed in *out-out* and *in-in* patterns, represented by flat curves in Fig. 6a. The results are from the 3-months dialogs dataset. The disassortative measures are already present in smaller-depth dialogs, although with a different negative slope. In Fig. 6b the results for the *in-out* strength mixing are compared for 2-months and 3-months dialogs networks. These findings suggest that a particular pattern with a large number of small-degree nodes communicating with one large-degree node occurs very often. This dynamical pattern, based on the friendship connections in MySpace which already shows some tendency towards disassortative mixing [30], is in clear contrast with the assortativity in common social structures [36] and static friendship networks in Facebook [4].

Confirmation of the weak-ties hypothesis. For quantitative measures of traditional social dynamics weak-ties hypothesis, it is widely accepted to determine correlations in *betweenness* centrality B_{ij} of a link and *overlap* O_{ij}

of two neighboring nodes on the network [37, 14]. Specifically, for the link (ij) , betweenness $B_{ij} = \sum_{s,t \neq i,j} \frac{\sigma_{st}(ij)}{\sigma_{st}}$ is given by the fraction of shortest paths $\sigma_{st}(ij)$ between pair of nodes (s,t) that pass through that link, compared to all shortest paths between (s,t) and averaged over all pairs on the network [32]. Note that for this purpose the network is considered as undirected! Overlap of two adjacent nodes i and j is computed as $O_{ij} = \frac{m_{ij}}{q_i + q_j - 2 - m_{ij}}$, where q_i and q_j stands for total degree of nodes i and j and m_{ij} is the number of nodes that are common neighbors to both of them. In traditional social dynamics it is expected that the overlap increases with bond strength, i.e., $O_{ij}(W) \sim W^{\eta_1}$. Moreover, due to the ubiquitous community structure, the bonds with large betweenness, e.g., connecting different communities, should not have large overlap. Consequently, $O_{ij}(B) \sim B^{-\eta_2}$. In the networks of online social contacts weak-ties hypothesis was confirmed in e-mail networks [37] and online games [14]. It was conjectured in [14] that universal exponents $\eta_1 = 1/3$ and $\eta_2 = 1/2$ should apply. Our results for MySpace two-months dialogs network and for more compact Net3321, shown in Fig. 7a,b, confirm this conjecture of Ref. [14]. In the insets we also show computed histograms of the betweenness $P(B)$ and of the weight $P(W)$ for all links in two-months dialogs window. Power-law tails in these distributions suggests diversity in both the organization of the communities and in the intensity of communications inside these communities.

4 Activity Patterns and Emotion Flow

Temporal patterns of user activity. In our high-resolution data information about every user activity over time can be presented as a pattern, an example is shown in lower panel in Fig. 8. Time in the original resolution is plotted along x-axis, and each integer number along y-axis stands for an user index. The indexes are ordered by user's first appearance in the dataset.

Two characteristic features of these temporal patterns are (cf. Figs. 8):

- *Arrival of new users*, depicted with the top boundary of the pattern, follows daily cycles. With the appearance of new users (relative to the beginning of the dataset) the system experiences increased activity, manifested as larger density of points below each “wave” of new users. Note also the stripes inclined upwards, which indicate possible correlated actions of the users involved in later times (in section 5 the temporal correlations are studied in detail). Both the arrival of new users and the increased activity of all other users obey periodicity, compatible with the *circadian cycles*, which is carried over from user's offline life. Similar features are observed in Blogs and Digs [17] and other social systems [7], confirming the importance of circadian cycles in the online dynamics.
- *Delay (interactivity) times of user actions Δt* , defined as time between two consecutive actions of a user, is a quantitative measure of fractality of the temporal activity pattern. Namely, following the line of a given user we find no characteristic distance between two consecutive activity points. The distribution of distances Δt between subsequent points for a given user, then averaged over all users in the dataset, is shown in Fig. 8 top panel. The broad distribution $P(\Delta t)$ shows faster decay, approximately as $\sim (\Delta t)^{-1.475 \pm 0.099}$, for the short delays in the range from $[5 - 85]$ minutes, the slope is indicated by dashed line. Whereas, a majority of the delay times appear to be one day, corresponding to the peak in the middle, or longer. These long delays can be fitted with the expression $P(\Delta t) = B(\Delta t)^{-\tau_\Delta} \exp\{-(\Delta t/\Delta_0)^\sigma\}$, with the parameters $\tau_\Delta = 1.061 \pm 0.009$ and $\Delta_0 = 75600$ and $\sigma = 2$. These parameters are for the 3-months window dataset, the fitted curve is shown by dotted line in Fig. 8 top panel. Same expression with a similar exponent, the cut-off $\Delta_0 = 52000$ and the stretching $\sigma = 3$ fits the dataset from 2-months time window, shown in Fig. 8 by dotted curve.

Occurrence of the power-law decay in the delay-time distribution was recently observed in different types of data related to online dynamics [5, 11, 38]. Theoretical arguments of the random queuing processes have been used to derive the universal scaling exponent $3/2$ in the delay-time distributions [39]. The present analysis of the data from MySpace social network suggests the exponent close to $3/2$ only for the short delay-times (between 5 and 85 minutes). However, the one day or longer delays are more probable with the exponent $\tau_\Delta \gtrsim 1$, suggesting another possible mechanism. It is also interesting to note that delays shorter than 5 minutes are equally probable, which suggests $t_{bin} = 5$ minutes as natural temporal resolution for these type of processes.

Structure of the emotional dialogs. We consider in detail the emotion content transmitted along the links in the network Net3321. As stated above, the directed weighted link W_{ij} on this network represents the number of messages sent from the node i to the node j within the time window of two months. Here we also include the emotion valence of these messages. By summing up the emotion contained in each message along the link we obtain the overall valence of the link, i.e., as positive, negative or neutral link. The network is shown in Fig. 9a with the links colored according to their emotion content—red (positive), black (negative) and blue (neutral). The size of the nodes indicates their degree centrality on the network. Each node carries a label—unique user ID index from the original data. The zoomed part of the network, which is displayed in Fig. 9b, illustrates a typical structure of the emotion-carrying links between hubs and the number of small-centrality nodes surrounding them.

Considering the emotion contents of the messages along the links, we find that the positive emotion (attractiveness) dominates the connections in MySpace dialogs. Whereas, the links carrying negative emotion (aversiveness) occur rather sporadically, acting as “impurity” in the sea of positively charged contacts. The temporal correlations of the positive (negative) emotion messages are further studied in section 5. Here we analyse the topology of the emotional (and otherwise important) connections on the network. We extract the subnetwork of negative links, whose fragments are found in different parts of the Net3321. The largest connected component of the negatively linked network is shown in Fig. 10a. Enlarged part of this network, shown in Fig. 10b, demonstrates typical flow patterns of the negative-emotion messages. Specifically, a node may act as a source or a sink of the negative links, can transmit or disseminate the emotion, or be involved in multiple reply-to events with the same emotion valence. Note that these two subnetworks with positive and negative emotion links are integrated into each other, and also that some messages (links) can be considered as neutral, i.e., carrying information, but not emotion. Therefore, the nodes that change the valence of the emotion messages exist and have a special role as pinning centers for the propagation of the emotional bursts on the network.

For most of the links the computed overlap is related with the widths of links, cf. Fig. 7 according to the social tie hypothesis. However, sometimes strong links, i.e., carrying large number of messages appear (visible on the network in Fig. 9) dis-proportionally to the topological centrality of the adjacent nodes. To find how the most of messages (and emotion) flow on the entire network, we analyse the maximum-flow spanning trees. These are suitable representation of the network where each node is connected to the tree by its *strongest link*. In Fig. 11 the maximum-flow spanning tree of the Net3321 is shown. The tree is constructed using a variant of greedy algorithm by ordering the *total weights* of the links $W_{ij} + W_{ji}$. It shows considerable side branching, which suggests heterogeneity in the intensity of the dialogs inside the existing communities. Moreover, it often occurs that a small-degree node interpolates in the branching process and transmits the flow of messages between the hubs. This feature is in agreement both with the observed community structure and the dis-assortativity of the dialogs-based network, discussed in section 3. It further supports the conclusion that the dynamical structure in the online social network MySpace is different from the networks of conventional social contacts.

5 Correlations in time series with emotional messages

The stochastic processes governing the communication with emotional messages among the users in MySpace can be studied from the point of view of the time-series analysis. From the datasets of the networked dialogs in a given time window various time-series are constructed here, for instance, the series which contain the *number of all identified messages per small time bin*. Similarly, we construct the time-series of the number of messages carrying positive/negative emotion valence. The time bin $t_{bin} = 5$ minutes is used as the natural resolution in these data. Examples of these time-series and their power spectra are shown in Fig. 12.

The time series in Fig. 12 exhibit fluctuations with strong daily periodicity (circadian cycles), observed also in the activity patterns in Fig. 8 and discussed above. This periodicity is manifested as a pronounced peak in the power spectrum at the corresponding frequency, i.e., at the index value $\nu \approx 56$ in this case. Apart from the peak, the power spectra in Fig. 12 are of the colored-noise type (fractal time series). Specifically, the spectrum can be expressed as $S(\nu) \sim 1/\nu^\phi$ for a range of frequency index below approximately 3000 (corresponding to the time scale longer than 2 hours) in the case of time series of all messages. Similar feature is found in the time series of messages with positive emotion for $\nu < 1000$ (or $t > 5$ hours) approximately. These features of the time series suggest occurrence of the long-range correlations in the fluctuations in number of messages of all types and in the messages with

positive-emotion valence. Whereas the spectrum of the negative-valence messages appears to be much closer to the white noise signal ($\phi \gtrsim 0$). The corresponding exponents are: $\phi^a = 0.59 \pm 0.08$, $\phi^+ = 0.55 \pm 0.08$, and $\phi^- = 0.15 \pm 0.06$ where symbols $a, +, -$ stand for all messages, and messages with positive and negative emotion valence, respectively. For completeness, we have also computed the Hurst exponent H , which measures the strength of fluctuations of these time series. In particular, for the time series of length T , it is determined from the power-law segment in the dependence $D(n)/\sigma(n) \sim n^H$, where $D(n)$ is the maximal deviations of the cumulative time series $\sum_k^n (Y(k) - \langle Y \rangle)$ and $\sigma(n)$ its standard deviations in a time window of varying length $n = 1, 2, \dots, T$. We find the following values $H^a = 0.82 \pm 0.03$, $H^+ = 0.83 \pm 0.04$, and $H^- = 0.62 \pm 0.03$ for these three time series. It is interesting to note that the Hurst exponents for all time series are larger than $1/2$, suggesting the *persistent fluctuations* in the overall activity and in the emotional messages of both polarity. Moreover, the scaling relation $\phi^a = 2H^a - 1$ is satisfied (within numerical error bars). Further study of the mechanisms behind these fractal time series and the avalanches of emotional comments is left out of this work [40].

6 Conclusions

We have performed a comprehensive analysis of the empirical data of user dialogs from the social network MySpace focusing on the quantitative analysis of users collective behavior. Our methodology, that can be used across large class of online systems with user-to-user interactions and high-resolution data, includes: compiling the data of suitable structure, extracting emotion content from texts of messages by automated methods tailored for this type of textual documents, and analysing the data with graph theory and statistical physics by properly accounting for the nonlinear dynamic effects. Our main findings can be summarized as follows:

- *Dialogs-based networks* in MySpace are dynamical structures built upon “friendship” connections. They organize in a large number of communities of various sizes and the “weak-tie” hypothesis holds in a manner similar to online games and e-mail networks. The actual use of links (within a given time window) reveals unbalanced message flow, yielding different organization of in-coming and out-going links; several hubs emerge to whom communication is directed from a large number of “small” nodes, manifested in strongly dis-assortative mixing; Furthermore, the emotion content of the messages passed along these links, averaged over the entire time window, is dominated by positive emotion (attractiveness), while the links with negative emotion (aversiveness) appears less often, but make a specific local pattern on the network.
- *Self-organized processes* of message exchange among the users have long-range temporal correlations of various degrees and persistent fluctuations, which clearly depend on the emotion content of the messages.
- *Robust patterns of user behaviors* are observed, which are linked with circadian cycles. Power-law distribution of delay-times with a small exponent for the delays longer than one day is found, suggesting that mechanisms different from queuing of tasks may be responsible.

In conclusion, the presented multi-analysis approach sheds a new light onto the actual problem of the functional structure in the online social networks. The studied patterns in MySpace, considered as one of the largest “social sites”, reveals dynamical structure that by many measures disagree with the common social networks. The self-organized dynamics with message exchange leads to specific organization of users where large diversity of groups is found as well as the importance of individuals within these groups. Moreover, the emerging collective behaviors depend on both intensity of communications and amount of emotion passed with the messages. Disproportional dominance of positive emotions (attractiveness) may also suggest presence of euphoria (“24 hours party”-like behavior). Our quantitative analysis with the results presented in sections 3, 4, and 5 can help in the ongoing social and psychology research on the problems “who” uses the social networks? and “how” the online social networks are used? and serve as a basis for further research and theoretical modeling.

Acknowledgments: Funding for this research was received in parts from the program P1-0044 of the Research agency of the Republic of Slovenia, the project no. P1-0044-3. M.Š. also thanks the research projects ON171037 and III41011 of the Republic of Serbia and the project from FP7-ICT-2008-3 under grant agreement n° 231323. We would like to thank George Paltoglou for providing user-friendly version of the emotion classifier described in Ref. [15].

Author's Contribution: Designed the research: BT; Contributed new software and compiled the data: MŠ; Performed emotion classification of the data: MM; Analysed the results and produced figures: BT, VG, MŠ, MM; Wrote the paper: BT.

References

- [1] Giles J. Social science lines up its biggest challenges. *Nature*. 2011;470:18–19.
- [2] Kleinberg J. The Convergence of Social and technological Networks. *Communications of the ACM*. 2008;51:66–72.
- [3] Cho A. Ourselves and Our Interactions: The Ultimate Physics Problem? *Science*. 2009;325.
- [4] Ugander J, Karrer B, Backstrom L, Marlow C. The Anatomy of the Facebook Social Graph. *arxiv:11114503v1*. 2011;.
- [5] Castellano C, Fortunato S, Loreto V. Statistical physics of social dynamics. *Reviews of Modern Physics*. 2009;81(2):591–646. Available from: <http://dx.doi.org/10.1103/RevModPhys.81.591>.
- [6] Guimerà R, Danon L, Díaz-Guilera A, Giralt F, Arenas A. Self-similar community structure in a network of human interactions. *Phys Rev E*. 2003 Dec;68:065103. Available from: <http://link.aps.org/doi/10.1103/PhysRevE.68.065103>.
- [7] Malmgren RD, Stouffer DB, Campanharo ASLO, Amaral LA. On Universality in Human Correspondence Activity. *Science*. 2009;325(5948):1696–1700. Available from: <http://dx.doi.org/10.1126/science.1174562>.
- [8] Johnson NF, Xu C, Zhao Z, Ducheneaut N, Yee N, Tita G, et al. Human group formation in online guilds and offline gangs driven by a common team dynamic. *Phys Rev E*. 2009 Jun;79:066117. Available from: <http://link.aps.org/doi/10.1103/PhysRevE.79.066117>.
- [9] Amichai-Hamburger Y, Vinitzky G. Social network use and personality. *Computers in Human Behavior*. 2010;26(6):1289 – 1295. Online Interactivity: Role of Technology in Behavior Change. Available from: <http://www.sciencedirect.com/science/article/pii/S0747563210000580>.
- [10] Panzarasa P, Opsahl T, Carley KM. Patterns and Dynamics of Users' Behavior and Interactions: network Analysis Off and Online Community. *Journal of the American Society for Information Science and Technology*. 2009;60:911–932.
- [11] Mitrović M, Tadić B. Bloggers Behavior and Emergent Communities in Blog Space. *Eur Phys Journal B*. 2010;73(2):293–301.
- [12] Szell M, Lambiotte R, Thurner S. Multirelational organization of large-scale social networks. *Proceedings of the National Academy of Sciences USA*. 2010;107(31):13636–13641. Available from: <http://www.pnas.org/content/107/31/13636>.
- [13] Mitrović M, Paltoglou G, Tadić B. Networks and emotion-driven user communities at popular blogs. *European Physical Journal B*. 2010 Oct;77:597–609.
- [14] Szell M, Thurner S. Measuring social dynamics in a massive multiplayer online game. *Social Networks* 2010; 39, 313–329.
- [15] Paltoglou G, Thelwall M, Buckley K. Online textual communication annotated with grades of emotion strength. In: *Proceedings of the Third International Workshop on EMOTION (satellite of LREC): Corpora for research on emotion and affect*; 2010. p. 25–30.
- [16] Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A. Sentiment strength detection in short informal text. *J Am Soc Inf Sci Technol*. 2010 December;61:2544–2558. Available from: <http://dx.doi.org/10.1002/asi.v61:12>.

- [17] Mitrović M, Paltoglou G, Tadić B. Quantitative analysis of bloggers' collective behavior powered by emotions. *Journal of Statistical Mechanics: Theory and Experiment*. 2011 February;2011(02):P02005+. Available from: <http://dx.doi.org/10.1088/1742-5468/2011/02/P02005>.
- [18] Dodds PS, Harris KD, Koloumann IM, Bliss CA, Danforth CM. Temporal patterns of happiness and information in a global social network: Hedonometric and Twitter. *arXiv:11015120v3*. 2011; Available from: <http://arxiv.org/abs/1101.5120v3>.
- [19] Chmiel A, Sobkowicz P, Sienkiewicz J, Paltoglou G, Buckley K, Thelwall M, et al. Negative emotions boost user activity at BBC forum. *Physica A*. 2011;390:29362944.
- [20] Coan JA, Allen JJB, editors. *The Handbook of Emotion Elicitation and Assessment*. Oxford University Press Series in Affective Science; 2007.
- [21] Scherer K. What are emotions? And how can they be measured? *Social Science Information*. 2005;44(4):695–729. Available from: <http://ssi.sagepub.com/cgi/content/abstract/44/4/695>.
- [22] Russell JA. A circumplex model of affect. *Journal of Personality and Social Psychology*. 1980;39:1161–1178. Available from: <http://dx.doi.org/doi/10.1037/h0077714>.
- [23] Calvo RA, D'Mello S. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *Affective Computing, IEEE Transactions on*. 2010 jan;1(1):18–37.
- [24] Bradley MM, Lang PJ. Affective norms for English words (ANEW): Instruction manual and affective ratings. The Center for Research in Psychophysiology, University of Florida.; 1999. Available from: <http://dionysus.psych.wisc.edu/methods/Stim/ANEW/ANEW.pdf>.
- [25] Paltoglou G, Gobron S, Skowron M, Thelwall M, Thalmann D. Sentiment analysis of informal textual communication in cyberspace. In: *Proc. ENGAGE 2010 (Springer LNCS State-of-the-Art Survey)*. Heidelberg: Springer; 2010. p. 13–23.
- [26] Ryan T, Xenos S. Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage. *Computers in Human Behavior*. 2011;.
- [27] Yarkoni T. Personality in 100.000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*. 2010;44:363–373.
- [28] Cheung CMK, Chiu PY, Lee MKO. Online social networks: Why do students use facebook? *Computers in Human Behavior*. 2011;27:1337 – 1343.
- [29] Ferrara E, Meo PD, Fiumara G, Provetti A. The role of strong and weak ties in Facebook: a community structure perspective. *Procedia Computer Science: International Conference on Computational Science, ICCS 2012*. 2012;p. 1–10.
- [30] Ahn YY, Han S, Kwak H, Moon S, Jeong H. Analysis of topological characteristics of huge online social networking services. *Proceedings of the 16th international conference on World Wide Web*. 2012;p. 835 – 844.
- [31] Fully anonymized data available upon registration from <http://www-fl.ijs.si/~tadic/Art/SI/>
- [32] Bollobas B. *Modern Graph Theory*. Springer-Verlag, Berlin Heidelberg; 1998.
- [33] Mitrović M, Tadić B. Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities. *Phys Rev E*. 2009 Aug;80(2):026123–+.
- [34] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*. 2008;2008(10):P10008 (12pp). Available from: <http://stacks.iop.org/1742-5468/2008/P10008>.

- [35] Garlaschelli D, Loffredo M. Patterns of link reciprocity in directed networks. *Phys Rev Lett.* 2004;93:268701.
- [36] Newman MEJ. Mixing patterns in networks. *Phys Rev E.* 2003;67:026126. Available from: <http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0209450>.
- [37] Onela J, Saramaki J, Hyvönen J, Szabo G, de Menezes M, Kaski K, et al. Analysis of large-scale weighted networks of one-to-one human communications. *New Journal of Physics.* 2007;9(6):176.
- [38] Vázquez A, Oliveira JG, Dezsö Z, Goh KI, Kondor I, Barabási AL. Modeling bursts and heavy tails in human dynamics. *Phys Rev E.* 2006 Mar;73(3):036127.
- [39] Grinstein G, Linsker R. Power-law and exponential tails in a stochastic priority-based model queue. *Phys Rev E.* 2008;77(1):012101–+.
- [40] Gligorićević V, Tadić B. Criticality of emotional avalanches in online social networks. in preparation. 2012;.

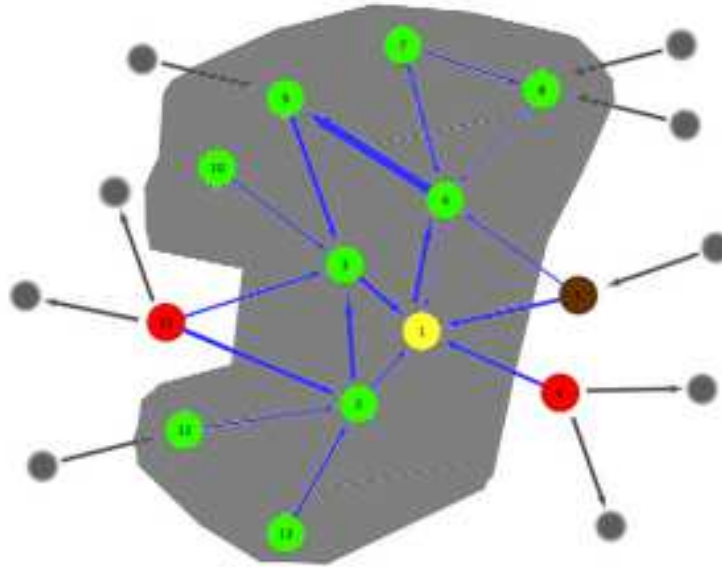


Figure 1: Schematic view of the parametrized breadth-first search of the dialogs occurring within specified time window, starting from a given user in MySpace, marked as node “1”. Red nodes “6” and “11” are examples of users who do not allow public access to their “walls”, however, their messages left on the neighboring “walls” can be identified. These nodes and their links are not included in the data. The presence of bots, companies, or another non-ordinary users is identified, an example depicted by the node “5”, and dropped from further search.

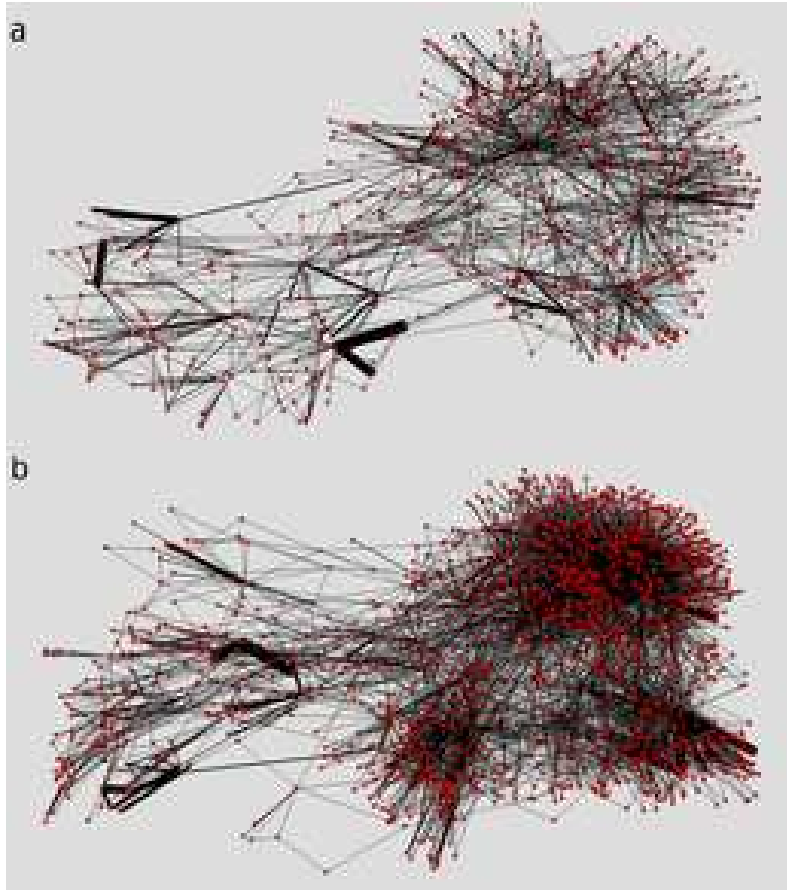


Figure 2: Initial part of the network of users connected by dialogs in *MySpace*, as compiled by our crawler for the time window of two months (a) and three months (b) starting from the same user-node.

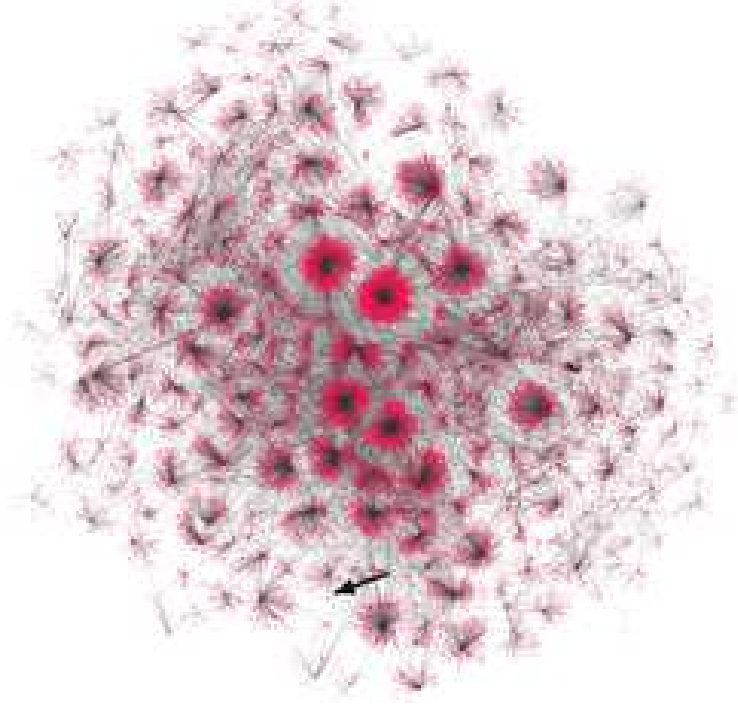


Figure 3: View of the network Net 2M from the dataset of dialogs with two months depth in MySpace. $N_U = 33649$ nodes are organized in 87 communities, seen as blobs of different sizes.

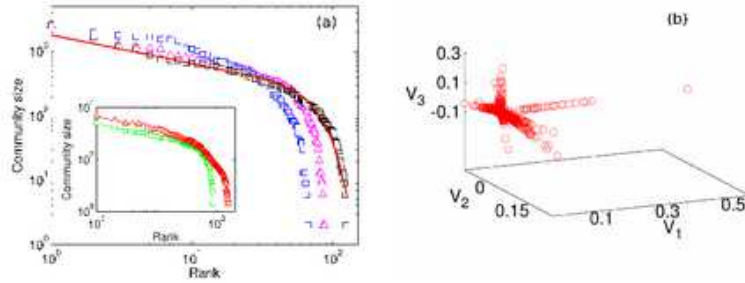


Figure 4: (a) Ranking distribution of the size of communities in the networks Net 2M of two months dialogs in MySpace. Three curves from front to back are for higher, standard, and lower resolution, respectively. Fit line has the slope 0.44 ± 0.02 and the cut-off 81.85 ± 0.78 . Inset: Same but for two and three months dialogs networks Net 2M (\square) and Net 3M (\triangle) with standard resolution. (b) The presence of communities in Net 3321 indicated by the branches in 3-dimensional plot of the eigenvectors for three lowest nonzero eigenvalues of the Laplacian.

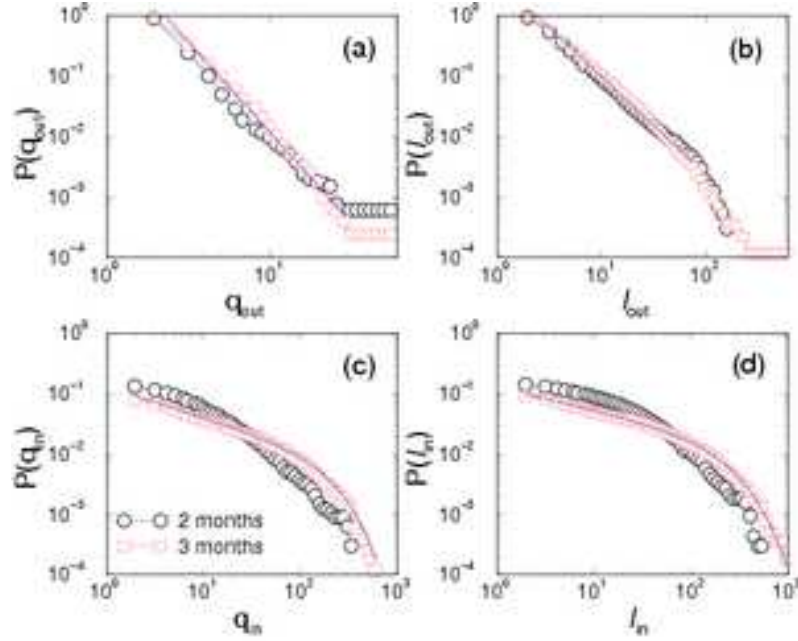


Figure 5: Out-degree (a) and Out-strength (b), and In-degree (c) and In-strengths (d) of nodes on the networks of MySpace dialogs within 2-months and 3-months time window. Log-binned data. Dotted lines are fits according to Eq. (1).

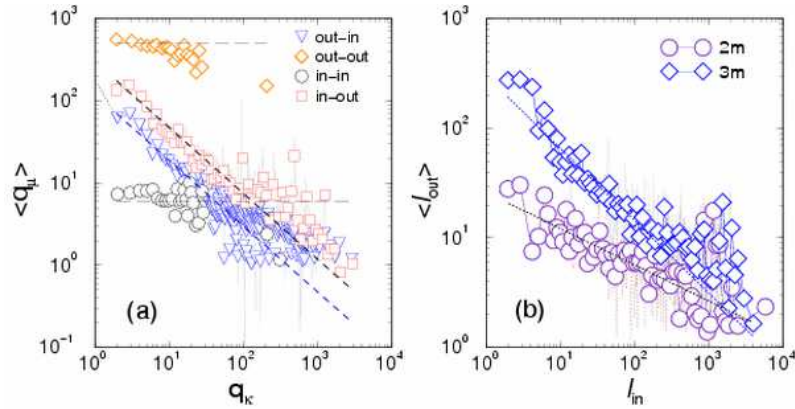


Figure 6: Mixing patterns on MySpace dialogs network: (a) correlations between in- and out-degree (four combinations) for the dialogs within 3 months time window, and (b) in- and out-strengths for 3 months and 2 months time window. Log-binned data. Dashed lines indicate slope $\mu \sim 1$, while two dotted lines are for $\mu = 0.33$ and 0.86 .

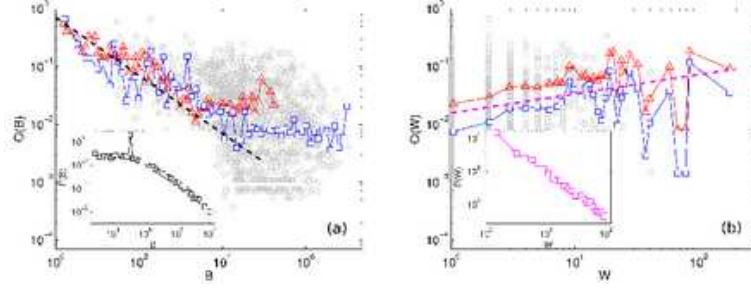


Figure 7: Averaged overlap as function of betweenness centrality B (a) and of weight W (b) for symmetrical links on MySpace dialog network of 2 months time window (\square) and Net 3321 (\triangle). Dashed lines in the left and the right panel indicate slopes $-1/2$ and $+1/3$, respectively, conjectured in Ref. [14]. Insets: Distributions $P(B)$ of betweenness and $P(W)$ of weights for the network of 2 months depth. Data are logarithmically binned.

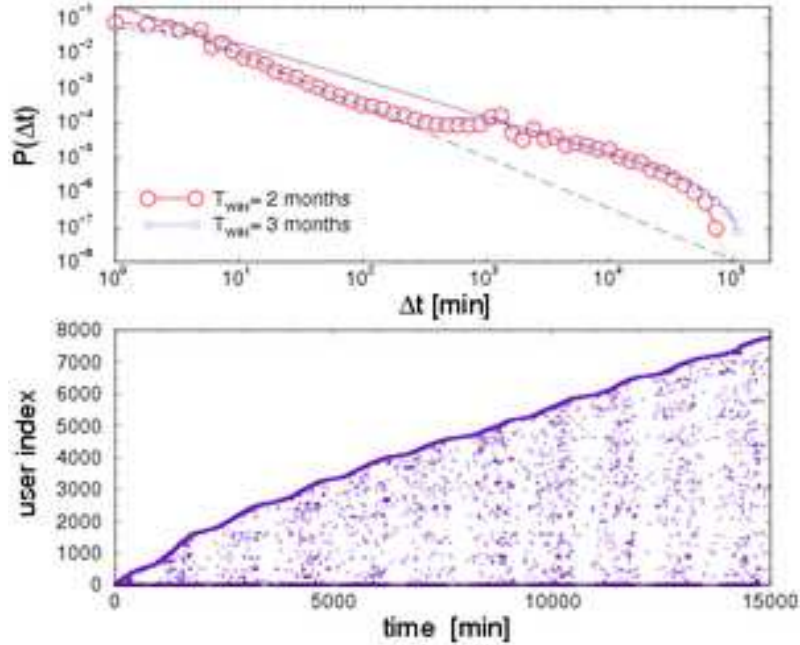


Figure 8: (bottom) Example of user-activity pattern from MySpace dialogs. Shown are first 10 days from the 2-months time window data. (top) Distribution $P(\Delta t)$ of the user delay time Δt , averaged over all users. Two curves are for the 2-months and 3-months time window datasets. Data are logarithmically binned. Dashed and dotted lines are explained in the text.

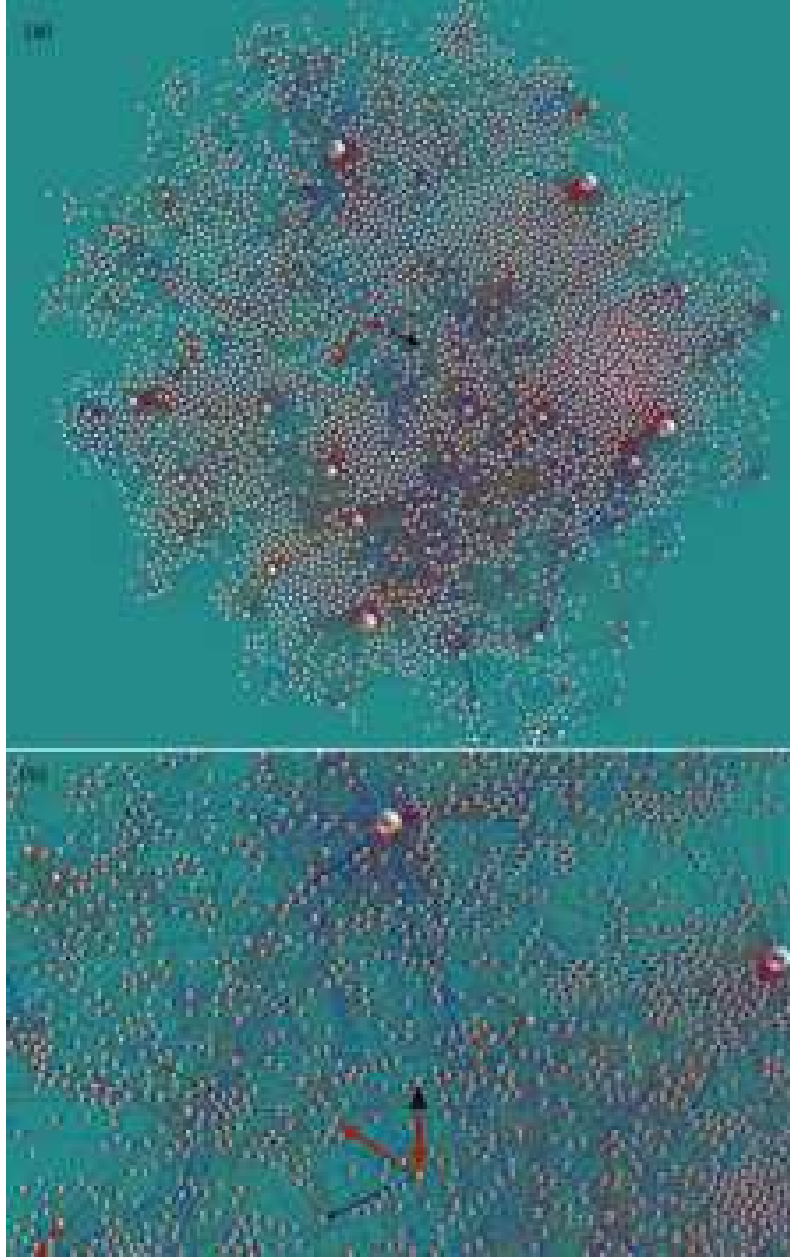


Figure 9: MySpace dialogs network Net3321 (a) and a zoomed part of it (b). Size of nodes corresponds to node's degree centrality. The width of links represents the cumulative number of messages within two months period, while the color indicates their emotion valence—positive (red), neutral (blue) and negative (black).

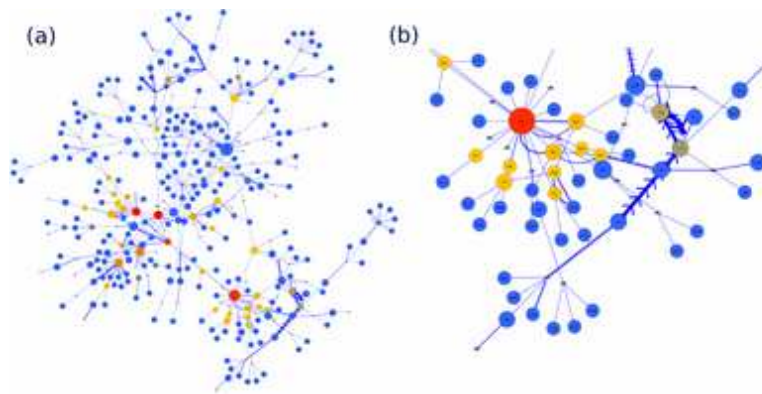


Figure 10: Largest connected component of the subnetwork with negative dialogs (a) and a its zoomed part (b), on which directions and weights of the negative valence dialogs are indicated. Nodes are marked by different size comparable with the out-degree and color determined by the betweenness-centrality of the nodes on the entire network Net3321, from which the negative links subnetwork is extracted.

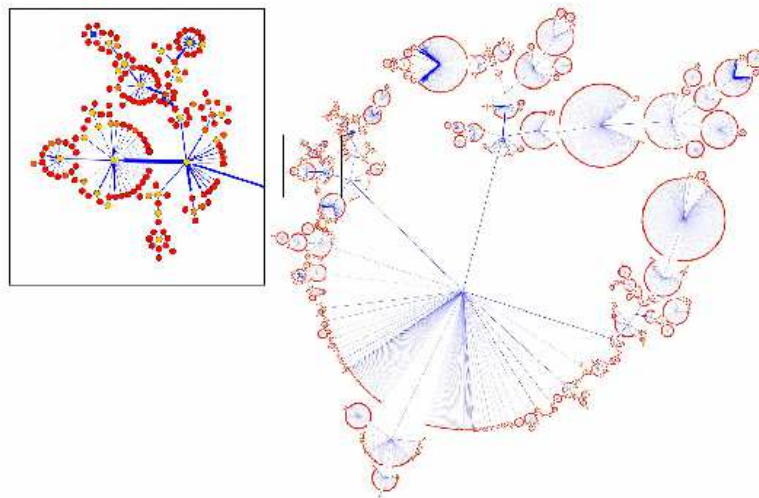


Figure 11: Maximum-flow spanning tree of the network Net3321. Enlarged part on the left shows a typical structure away from the root.

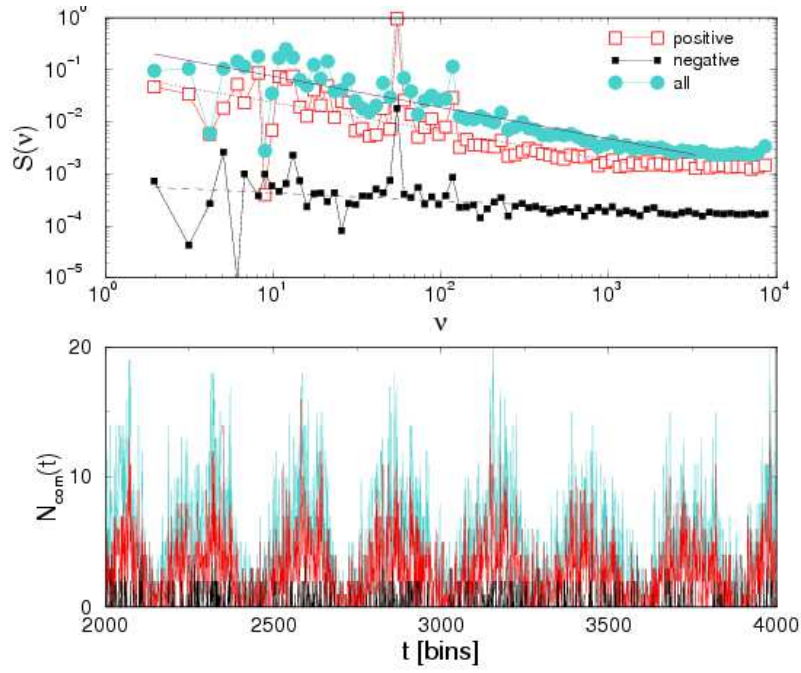


Figure 12: Time series of the number of messages from MySpace dialogs network for 2-months time window: all identified messages (cyan), and the messages classified as carrying positive (red) and negative (black) emotion valence. Time axis in bins of 5 minutes. Length of each time series is 16384 time bins. For better vision shown is only a small part corresponding to one week time span. Top panel: power spectrum of these time series. Log-binned data. Fit lines are explained in the text.